

# Sampling through time and phylodynamic inference with coalescent and birth-death models

Erik M Volz<sup>1\*</sup> and Simon DW Frost<sup>2</sup>

1. Department of Infectious Disease Epidemiology, Imperial College London, United Kingdom

2. Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom

\* Corresponding author: e.volz@imperial.ac.uk

August 29, 2014

## Abstract

Many population genetic models have been developed for the purpose of inferring population size and growth rates from random samples of genetic data. We examine two popular approaches to this problem, the coalescent and the birth-death-sampling model, in the context of estimating population size and birth rates in a population growing exponentially according to the birth-death branching process. For sequences sampled at a single time, we found the coalescent and the birth-death-sampling model gave virtually indistinguishable results in terms of the growth rates and fraction of the population sampled, even when sampling from a small population. For sequences sampled at multiple time points, we find that the birth-death model estimators are subject to large bias if the sampling process is misspecified. Since birth-death-sampling models incorporate a model of the sampling process, we show how much of the statistical power of birth-death-sampling models arises from the sequence of sample times and not from the genealogical tree. This motivates the development of a new coalescent estimator, which is augmented with a model of the known sampling process and is potentially more precise than the coalescent that does not use sample time information.

The genetic diversity of many pathogens is influenced by recent epidemiological history, and a variety of methods exist to estimate features of an epidemic history given random samples of pathogen genetic markers [1]. An issue that is central to how pathogen genetic diversity is understood is how infected individuals are sampled. A great deal of theory has been developed under the assumption of complete sampling, that is, that all infected individuals in the population are sampled and provide at least one pathogen sequence. These methods have found great utility for the study of small outbreaks [2, 3], and for certain hospital acquired infections [4]. A separate body of theory has developed for the study of epidemics where a sample of hosts is obtained for pathogen sequencing, and these methods are derived from classical population genetic models such as the coalescent [5, 6] and classical population dynamics models such as the birth death process [7, 8]. This manuscript considers the scenario of incomplete sampling and the potentially confounding effects of non-random sampling through time on inference using the coalescent and birth-death-sampling formula [9].

The coalescent is a mathematical model of genealogies and describes the structure of genealogies generated by different demographic processes [10]. The coalescent has been the standard tool for demographic inference and is the underlying genealogical model in most phylogenetic software [11, 12]. Under the neutral coalescent, the time between consecutive common ancestry events (the internode intervals) are modeled as a point process with a hazard rate  $r(t)$  that depends on

the effective population size  $N_e(t)$  and the number of extant lineages in that interval  $A(t)$  at time  $t$  in the past. With time in units of the generation interval  $\tau$ , this becomes

$$r(t) = \binom{A(t)}{2} / N_e(t)$$

By relating the time of common ancestry to the population size, the coalescent enables estimation of the latter. A variety of non-parametric [13, 14, 15] and parametric [16, 13, 17] models have been developed for  $N_e$  as a function of time. The parametric models for  $N_e(t)$  tend to be deterministic functions of time, and we will consider such deterministic models in this paper, although there have been several recent attempts to fit stochastic demographic process models using the coalescent [18, 19].

Birth-death processes trace their origins to work by Kendall[8], who showed how to calculate the probability that a given number of lineages will survive up to a given point of time in a stochastically growing population. Further results were developed by [20, 21], who showed how to calculate the probability density of genealogies generated by the birth-death process under complete sampling. These models were subsequently extended to account for incomplete sampling of the population by Stadler *et al.* [9]. In order to account for incomplete sampling, the birth-death process must be combined with a model of the sampling process. Two sampling processes have thus far been considered in birth-death-sampling models: sampling of lineages may take place at a constant rate; or, at a given point in time, a proportion of lineages may be sampled uniformly at random. These sampling processes may be combined, and recently-developed methods allow sampling rates to vary through time according to a step function [22].

There are many variations on the coalescent and birth death models that could be compared. Different coalescent and birth-death models make different assumptions about the demographic and sampling process, and each will be susceptible to different levels of bias by violation of those assumptions. We will focus on two models that have recently received considerable attention and have been used in epidemiological investigations. We use the birth-death-sampling model (henceforth abbreviated BDM) described in [9], and the the coalescent model (henceforth abbreviated CoM) for an unstructured population as described in [17]. Originally, CoMs were based on restrictive assumptions about the proportion of the population sampled and when taxa are sampled. CoMs were also based on strictly deterministic demographic processes, but all of these assumptions have been relaxed since the coalescent was first introduced. BDMs were originally based on census sampling at a single point in time, but that assumption has also been relaxed. Both models have been extended to consider heterogeneous structured populations [17, 23].

The likelihood of a genealogy given a demographic history may be calculated using either the CoM or the BDM, though these two models have very different mathematical foundations. The likelihood functions provided by each approach are difficult to reconcile mathematically, yet they tend to give similar results as we demonstrate below. The BDM has the advantage of accounting for stochasticity of the demographic process in an efficient and natural way. It is also possible to account for stochastically varying effective population size in the coalescent, but this has greater computational requirements [18]. A potential disadvantage of BDMs is that they require a model of the sampling process, whereas the coalescent makes no assumptions about how lineages are sampled through time. If the sampling process deviates from the simplistic processes that form the basis of current BDM theory, it is possible that estimates based on the BDM will be biased.

Both methods have particular advantages and vulnerabilities. Estimates based on CoMs may be biased by noisy demographic processes, and estimates based on the BDM may be biased by misspecification of the sampling process. In this manuscript, we will evaluate the vulnerability of

both methods to these confounders. Because of the additional assumptions about sampling built into BDMs, it is difficult to make a direct comparison of the statistical power of BDMs and CoMs. If the sampling process is correctly specified, the observed sequence of sample times provides a great deal of information about the population size through time, which is not directly accessible with the CoM approach. Indeed, given a sequence of sample times, it is possible to estimate birth and death rates without a genealogy provided that the model of the sampling process is correctly specified (section 2.1). We show that much of the statistical power of the BDM approach is derived from information in the sequence of sample times, and not in the genealogy. This finding also suggests an enhancement to CoMs: if the sampling process is known, we can augment the CoM likelihood with a separate likelihood for the sequence of sample times. This augmented coalescent method is presented in section 2.2.

If sampling at a single time point (homochronously) we show that estimates based on CoMs and BDMs are very similar. In section 3.7, we show how the distribution of coalescent times predicted by CoM converges with large sample size to the distribution given by BDM.

## 1 The demographic and sampling processes

The population size  $Y(t)$  is modelled as a continuous time Markov chain on  $[0, \infty)$ , which is governed by the following transition probabilities:

$$\begin{aligned} P(Y(t + \Delta t) = Y(t) + 1) &= \lambda Y(t) \Delta t + O((\Delta t)^2) \\ P(Y(t + \Delta t) = Y(t) - 1) &= \mu Y(t) \Delta t + O((\Delta t)^2) \\ P(Y(t + \Delta t) = Y(t)) &= 1 - (\lambda + \mu) Y(t) \Delta t + O((\Delta t)^2) \end{aligned} \tag{1}$$

where  $\lambda$  and  $\mu$  are the per-capita birth and death rates of the process, respectively. Initially,  $Y(0) = 1$ .

We investigated three distinct sampling processes for the reconstruction of genealogies from a simulated demographic history:

1. Continuous sampling through time at constant rate. According to this model, after a lineage dies (with a per-lineage rate  $\mu$ ), it is sampled with independent probability  $p$ .
2. Homochronous sampling. According to this model, every extant lineage at a predetermined time point is sampled with independent probability  $\psi$ .
3. Weighted sampling through time at changing rate. According to this model, a weighted sample of  $n$  lineages at the time of death is taken with sampling weights that depend on time. If  $\{t_i\}$  is the set of death times for lineages indexed by  $i$ , the sample weights are  $w_i = e^{\alpha t_i}$ .

Note that with the exception of homochronous sampling, the lineages are only sampled at the time of death. This design is chosen for mathematical convenience, since it eliminates the possibility of a zero-branch length in the genealogy, but both BDMs and CoMs can be generalized to this situation.

In this manuscript, we use CoMs based on the following deterministic approximation  $y(t)$  to the stochastic process  $Y(t)$ :

$$y(t) = y(0)e^{t(\lambda - \mu)}, \tag{2}$$

with real-valued initial conditions  $y(0)$  that will be estimated.

Genealogies were generated by simulation of the birth-death process in continuous time using the software MASTER 1.7.1 [24]. For simulating genealogies with a time-dependent sampling rate, we developed a custom simulator for the birth-death process in Python (see supporting information). We simulated 300 genealogies for each of the three sampling scenarios given above using  $\mu = 1$  and  $\lambda = 2$  or  $\lambda = 1.25$ . In the case of sampling through time, we terminated the simulation when 100 samples were collected and using a sampling probability of  $p = 1\%$  or  $p = 50\%$ . If sampling homochronously, we sampled 100 taxa after 9.21 or 25 units of time, yielding a sample proportion that varied around 1% or 20%, respectively. Simulations that failed to reach the target sample size were removed.

## 2 Estimation methods

All models are fitted by maximum likelihood (ML). The choice of ML was motivated by the simplicity of the demographic process, the small number of free parameters, and the possibility that arbitrary priors in a Bayesian framework might obfuscate the important differences between methods. For the exponential growth process, there are four potential parameters that could be estimated: birth rates  $\lambda$ , death rates  $\mu$ , the initial population size  $y(0)$  (needed for CoMs but not BDMs), and a parameter that describes sampling (needed for BDMs but not CoMs). As previously shown in the analysis of BDMs, at most two of these parameters are identifiable from a genealogy alone, and we must therefore choose which parameters to fix according to prior knowledge, and CoMs are subject to the same identifiability constraints. We focus on an epidemiologically plausible scenario, where birth rates and the number of infections are unknown, but independent clinical information provides information on death rates. Consequently, we will assume  $\mu = 1$  is known, and will focus on the estimation of birth rate  $\lambda$  along with the nuisance parameters describing initial population size (for CoMs) or sampling rates (for BDMs). We will also consider the special case of homochronous sampling, in which we can reparameterise the CoM such that, like the BDM, estimates of the sampling fraction can be obtained.

Throughout the remainder of the paper, we will use two symbols to denote time on different axes, and all dynamic variables will be defined on both axes.  $t$  will denote time from an arbitrary point in the past, while  $s$  will denote time before present. It will be useful to define the population genetic models in terms of the retrospective time axis  $s$ .

Let  $\mathcal{G} = (\mathcal{N}, \mathcal{E}, X)$  represent a genealogy consisting of a set of nodes  $\mathcal{N}$ , edges  $\mathcal{E}$  and a function  $X : \mathcal{N} \rightarrow \mathbb{R}$  that gives the time  $s$  before the present of each node. Every edge corresponds to a 2-tuple  $(u, v)$  such that  $u, v \in \mathcal{N}$  and the node  $u$  is said to be ancestral to  $v$ . We will consider only rooted binary genealogies; every internal node has exactly two descendents, and all internal nodes but the root have exactly one ancestor.

For CoMs, we use the likelihood given in [17], and we denote the MLE birth rate  $\hat{\lambda}^C$ . This likelihood is that of a time-inhomogenous point process with a hazard rate that depends on the population size and number of extant lineages. Specifically, following the approach in [17], the total population birth rate will be denoted  $f(s) = \lambda y(s)$  and the rate of coalescence is

$$\begin{aligned} r(s) &= f(s) \frac{\binom{A(s)}{2}}{\binom{y(s)}{2}} \\ &= \binom{A(s)}{2} \frac{2\lambda}{y(s) - 1}, \end{aligned} \tag{3}$$

where the term on the right can be understood as the hypergeometric probability of selecting two lineages that are ancestral to the sample out of the set of  $y(s)$  lineages. Now let  $x'$  denote the vector of node times (including sampled tips) in ascending order. The likelihood of the  $i$ 'th interval is

$$L_i = \begin{cases} e^{-\int_{x_i}^{x_{i+1}} r(s)ds} & x_{i+1} \text{ is a sample time} \\ r(x_{i+1})e^{-\int_{x_i}^{x_{i+1}} r(s)ds} & x_{i+1} \text{ is a coalescent time} \end{cases} \quad (4)$$

And the likelihood is

$$\mathcal{L}_{com}(\mathcal{G}|\lambda, \mu, y_0) = \prod_{i=0}^{2n-2} L_i \quad (5)$$

Note that the number of terms in the likelihood is the number of internode intervals  $2n - 2$  if all sampling times are distinct. One subtlety arises if more than one lineage is sampled at a single time point, such as with a homochronous sample, in which case we simply deduplicate elements in the vector  $x'$  and adjust the number of terms in the likelihood.

For BDMs, we used the ML framework described in [9]. We denote the MLE birth rate  $\hat{\lambda}^{BD}$ . The R package *expoTree* [25] was used along with the implementation described here, and all results presented below are based on the best performing of the two implementations of the BDM likelihood. We simplified the likelihood equations in [9] to two situations: sampling occurs at a single time point with sample fraction  $\rho$ , or individuals are sampled with probability  $p$  at the time of death. Let  $x$  denote the vector of times before present for each internal node in  $\mathcal{G}$  in descending order. Note that  $x_0$  corresponds to the root of the tree. If the sampling takes place according to the homochronous process,  $\rho$  will denote the probability of sampling a lineage at a single point in time. Then,

$$\mathcal{L}_{bdm}(\mathcal{G}|\lambda, \mu, \rho) = \lambda^{n-1} (4\rho)^n \prod_{i=0}^{n-2} q(x_i, c_2)^{-1} \left( \int_{x_{or}=x_0}^{\infty} q(x_{or}, c_2)^{-1} dx_{or} \right), \quad (6)$$

where  $q(\cdot)$  is derived from the birth-death sampling formula:

$$q(s, c) = 2(1 - c^2) + e^{-c_1 s}(1 - c)^2 + e^{c_1 s}(1 + c)^2, \quad (7)$$

and  $c_1$  and  $c_2$  are the following constants:

$$c_1 = |\lambda - \delta| \quad (8)$$

$$c_2 = -(\lambda - \delta - 2\lambda\rho)/c_1 \quad (9)$$

Note that the integral in the likelihood equation is an attempt to account for the unobserved time of origin of the birth-death process.

If sampling heterochronously at rate  $\psi = \mu p$ , the likelihood has a different form. Let  $x$  denote the vector times before present of each node as above, and let  $y$  denote the vector of sample times in any order.

$$\begin{aligned} \mathcal{L}_{bdm}(\mathcal{G}|\lambda, \mu, \psi) = & \int_{x_{or}=x_0}^{\infty} \lambda^{n-1} \psi^n \{ q(x_{or}, c_2)^{-1} \prod_{i=0}^{n-1} q(y_i, c_2) \prod_{i=0}^{n-2} q(x_i, c_2)^{-1} \\ & - q(x_{or}, c_3)^{-1} \prod_{i=0}^{n-1} q(y_i, c_3) \prod_{i=0}^{n-2} q(x_i, c_3)^{-1} \} dx_{or}, \end{aligned} \quad (10)$$

and,  $c_1$  and  $c_2$  and  $c_3$  are the following constants:

$$c_1 = \sqrt{(\lambda - \mu)^2 + 4\lambda\psi} \quad (11)$$

$$c_2 = (\mu - \lambda)/c_1 \quad (12)$$

$$c_3 = (\lambda + \mu)/c_1 \quad (13)$$

BDMs and CoMs were fitted according to the same numerical algorithm, with maximization of the likelihood accomplished in R using the simplex method. In order to assure convergence to the global maximum, multiple starting conditions were drawn from a multivariate uniform distribution, and the likelihood optimized for each. The best model fit is reported among the three or five optimizations, although in general they converged to the same value.

## 2.1 Estimating birth rates using times of sampling

Consider the sequence of sample times in increasing order  $\vec{t} = (t_1, \dots, t_n)$ . If the sampling process is known, the sequence of sample times are informative about population size. We will consider a simplistic sampling process such that individuals are sampled at a constant rate upon death, which is the sampling process underlying current BDMs. If sampling occurs at a constant known rate, it is straightforward to estimate the historical population size from the sample times, since the probability that a sample will be observed at some point in time is proportional to population size at that time. Therefore, it is possible to estimate the population size using sample time information alone, and not using the genealogy. We show here that it is possible to estimate the birth rate, even if the sampling rate is unknown. Two simple estimators for are presented. The first is based on a simple regression with the expected cumulative number of samples through time. The second is based on treating the sample times as arising from a point process and using maximum likelihood.

Let  $S(t)$  denote the cumulative number of samples collected up to time  $t$ . We show that the cumulative number of samples increase at the same rate as the unknown population size. According to the deterministic model, the expected change in  $S$  over time  $\Delta t$  will be

$$\begin{aligned} \Delta S(t) &= (\Delta t)p\mu y(t) + O((\Delta t)^2) \\ &= (\Delta t)p\mu y(0)e^{(\lambda-\mu)t} + O((\Delta t)^2) \end{aligned} \quad (14)$$

Consequently, the logarithm of  $S(t) \propto (\lambda - \mu)t$ . Regressing the vector  $\log(S(\vec{s}))$  on the vector  $\vec{s}$  yields an estimate of the growth rate  $k = \lambda - \mu$ , and using knowledge of  $\mu = 1$  we have the regression estimator

$$\hat{\lambda}^R = \hat{k} + \mu \quad (15)$$

Figure 1 shows the number of samples through time, for a single simulated genealogy along with the regression line.

The likelihood approach is based on modeling the sequence of sample times as a point process and also makes use of the deterministic approximation to population size. The rate of a sample appearing at time  $t$  is

$$f(t) = p\mu y(0)e^{(\lambda-\mu)t} = ae^{kt},$$

with  $a = p\mu y(0)$  and  $k = \lambda - \mu$ . The probability of  $\vec{s}$  is

$$P(\vec{t}|a, k) = \prod_{i=1}^n f(t_i) e^{-\int_{t_{i-1}}^{t_i} f(\tau) d\tau},$$

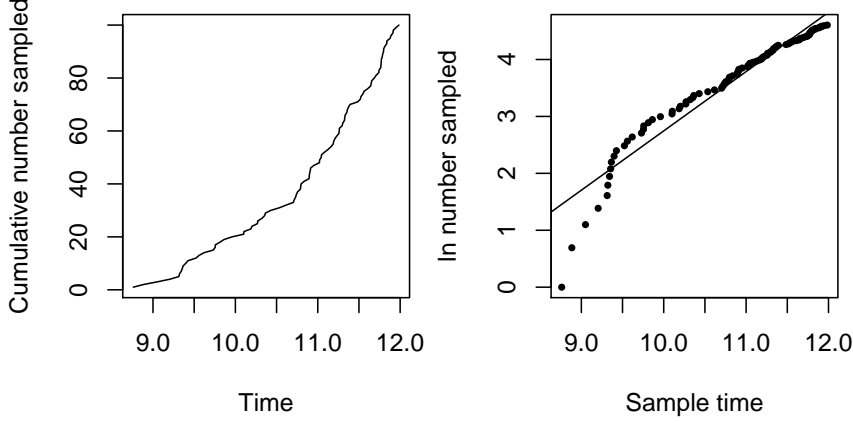


Figure 1: Left: Cumulative number of samples through time. Right: Log cumulative samples with regression line.

with  $s_0 = 0$  for consistency. As with the regression estimator,

$$\hat{\lambda}^S = \hat{k} + \mu \quad (16)$$

## 2.2 The augmented coalescent model

The genealogy  $\mathcal{G}$  and the sample times  $\vec{s}$  are conditionally independent given demographic and sampling parameters  $\theta = (\lambda, \mu, p, y(0))$ . Therefore, the likelihood of both is the product of the marginal likelihoods given above:

$$P(\mathcal{G}, \vec{t} | \theta) = P(\mathcal{G} | \theta) P(\vec{t} | \theta) \quad (17)$$

We will denote the MLE birth rate as  $\hat{\lambda}^A$ .

## 3 Results

The following results demonstrate the level of bias, precision and efficiency of different inference methods when estimating the birth rate from genealogies generated by the birth death process.

### 3.1 Constant sampling rate

Figure 2 shows the distribution of MLEs for five estimators presented above based on 300 simulated genealogies. Simulations were based on a sampling process with constant sampling probability  $p = 1\%$  at the time of death.

Estimators based only on the sequence of sample times  $\vec{s}$  perform well even though they do not use the coalescence times. The ML estimator  $\hat{\lambda}^S$  consistently outperforms the simple regression estimator  $\hat{\lambda}^R$ , presumably reflecting that residuals in the loglinear regression model are not i.i.d. normal.

Comparing a model that only uses sample time information ( $\lambda^S$ ) with the CoM that only uses genealogical information ( $\lambda^C$ ) we find that the RMSE of  $\hat{\lambda}^S$  is 11.5% compared to 11.8% for  $\hat{\lambda}^C$ . In this instance, the sample time sequence is actually more informative than the coalescent times for inferring birth rates.

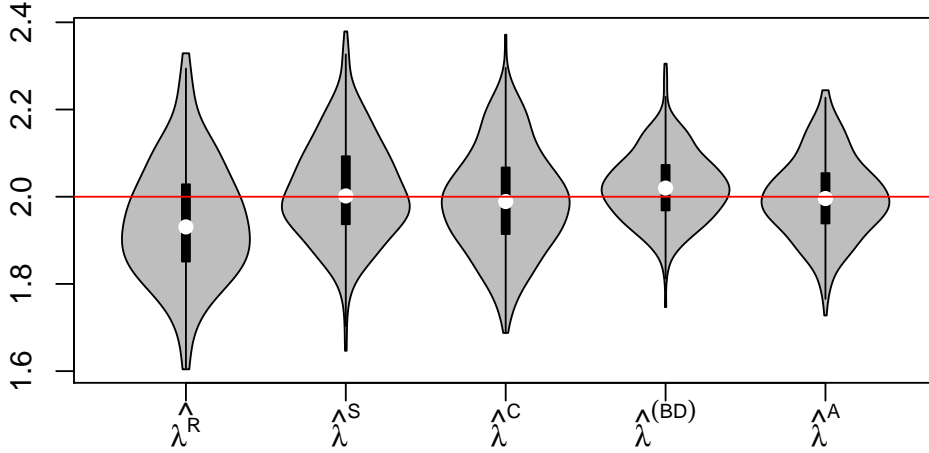


Figure 2: Distribution of MLE birth rates from 300 simulations with constant sampling rate and using five different estimators. The red line shows the true parameter value.  $\hat{\lambda}^R$  is a regression estimator,  $\hat{\lambda}^S$  is a ML estimator using sample times,  $\hat{\lambda}^C$  is the coalescent estimator,  $\hat{\lambda}^{(BD)}$  is the BDM estimator, and  $\hat{\lambda}^A$  is the coalescent estimator that also uses sample times.

Comparing the BDM and coalescent, we find that the BDM is more precise (RMSE = 0.085) but slightly less accurate; the average bias of the BDM estimator was 0.022 (95% CI:(0.013, 0.031) ) compared to 0.009(95% CI: (-0.022, 0.005)) for the CoM estimator. Comparing the BDM and augmented CoM (a model that uses both coalescence and sample times), we find that the augmented CoM is slightly less precise than the BDM (RMSE of  $\hat{\lambda}^A$  is 0.092), which may reflect the use of a misspecified deterministic population size, however, we did not detect significant bias of  $\hat{\lambda}^A$  (95% CI:(-0.012,0.009)) in contrast to the BDM.

Figure 3 sheds some light on why the estimators perform differently by comparing the maximum likelihood estimated by each method on each simulated genealogy. SI figure S1 shows a similar scatter plot of MLE birth rates. The BDM likelihood is highly correlated with that of all other estimators. In contrast, the CoM likelihood is almost independent of the estimators that use sample times only ( Pearson correlation = 0.066). The highest correlation is found between the BDM and the augmented CoM (Pearson correlation = 0.95). This illustrates that the CoM is not using sample time information, but the BDM and augmented CoM are using information from both the sample times and genealogy.

### 3.2 Homochronous sampling

If all samples are collected at a single point in time, and if the sampling proportion is unknown, then the time of sampling and sample size confer no information about population size. The homochronous sampling case with unknown sampling rate therefore provides a fair comparison for BDMs and CoMs. Here we consider 300 simulations of the birth death process with a sample of  $n = 100$  at  $t = 9.2$ , so that the sample fraction is around 1%, though it differs between replicates. The birth rate used in the simulations was  $\lambda = 2$ .

Figure 4 shows the distribution of MLE birth rates. The distributions are very similar and have similar precision (RMSE of  $\hat{\lambda}^{BD}$  is 0.106 and RMSE of  $\hat{\lambda}^C$  is 0.101). The CoM estimator



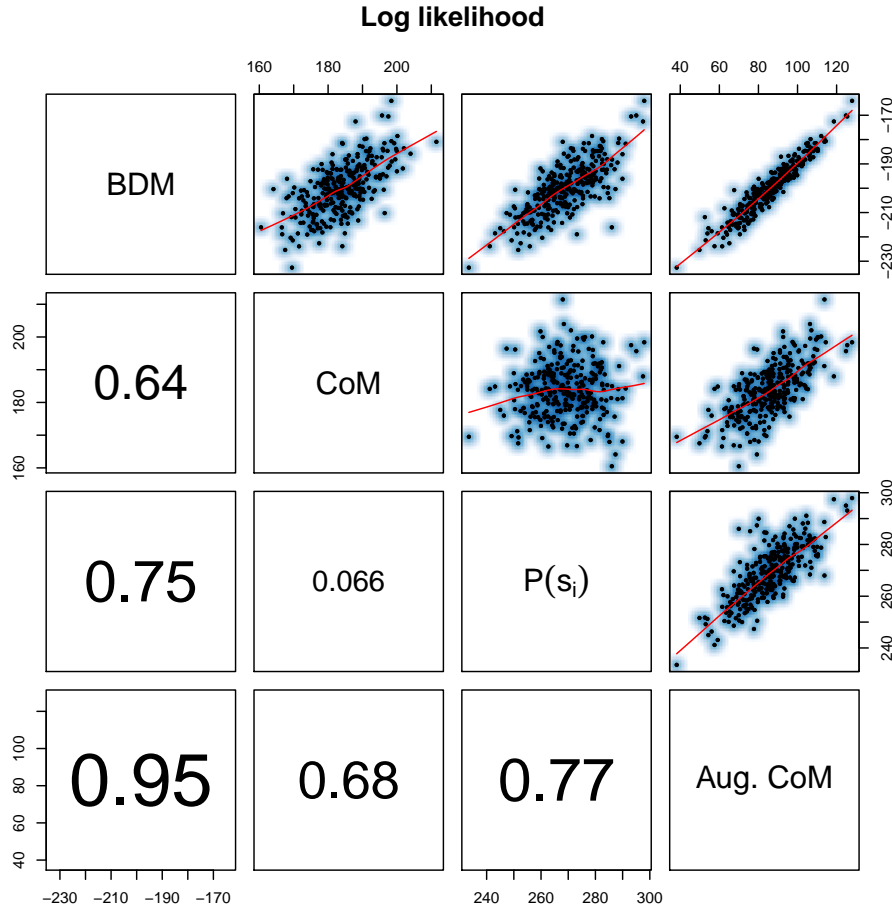


Figure 3: The log likelihood corresponding to MLEs from four estimation methods and based on 300 simulated genealogies. The Pearson correlation coefficient between log likelihoods is shown in the lower panels. The upper panels show a scatter plot with smoothing splines.

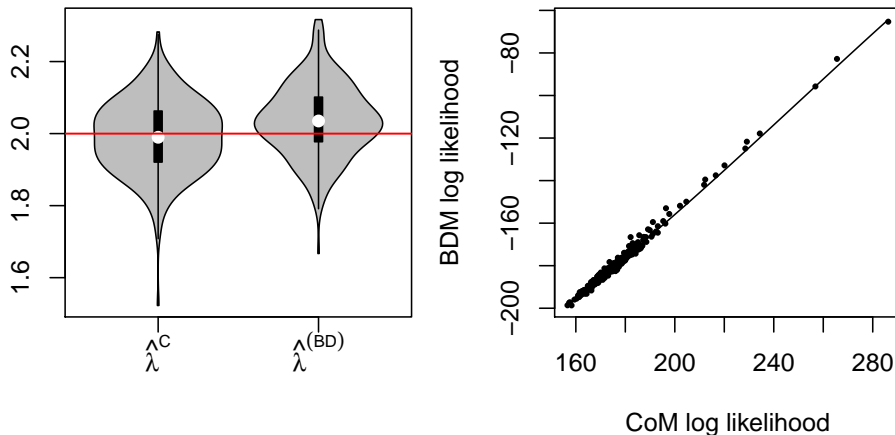


Figure 4: MLEs based on simulations with homochronous sampling. Left: Distribution of MLE birth rates from 300 simulations and using BDM and CoM estimators. The red line shows the true parameter value. Right: The log likelihoods of the BDM and CoM MLEs with a smoothing spline.

does not have detectable bias (95% CI of bias:  $(-0.0183, 0.0048)$ ), but the BD model slightly overestimates birth rates (average bias = 0.036, 95% CI:  $(0.0242, 0.0470)$ ). Figure 4 also shows that the log likelihoods of the MLEs generated by both methods are highly concordant up to a constant factor. The Pearson correlation of BDM and CoM maximum likelihoods is 99.6%. The estimated birth rates also have a high correlation coefficient of 86.6%.

Comparing the RMSE of the BDM estimator in both the homochronous and constant sampling rate cases, it appears that having informative sample time information decreases the residual sums-of-squares of the BDM estimator by about 36%, but this gain in precision will certainly depend on parameters of the system and sample size.

We repeated the simulation exercise with a smaller birth rate ( $\lambda = 1.25$ ) in order to assess if the CoM estimator would be less accurate if the population is growing more slowly. The MLEs are depicted in SI figure S2. With the smaller birth rate, we do not detect significant bias of the BDM estimator (average bias  $< 1e-3$ , 95% CI:  $(-0.0069, 0.0077)$ ), or with the CoM estimator (average bias 0.002, 95% CI of bias:  $(-0.0057, 0.0096)$ ). The RMSE of the BDM and CoM estimators are similar (0.037 and 0.039 respectively).

### 3.3 Comparison of estimated sample rates

An alternative parameterization of the coalescent is in terms of the population size at the time of sampling in a homochronous scenario. In this case, we can calculate a deterministic approximation to the population size at time  $s$  in the past as

$$y(s) = \frac{n}{\rho} e^{-s(\lambda - \mu)},$$

where  $n$  is the sample size and  $\rho$  is the sample proportion, and  $n/\rho$  is the population size at the time of sampling. According to this parameterization, we replace the nuisance parameter  $y(0)$  with  $\rho$ , and the coalescent estimates of the sample proportion can be directly compared to estimates with the BDM.

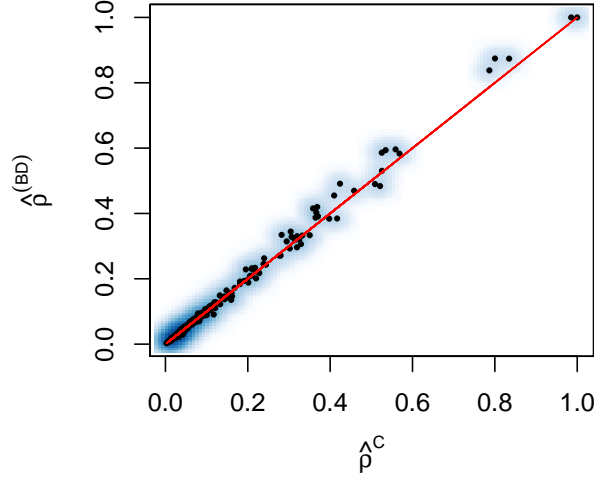


Figure 5: Estimated sample proportions using the coalescent and BDM with homochronous sampling.  $\lambda = 1.25, \mu = 1, n = 100$ .

We fit the reparameterized CoM to the the same genealogies used in section 3.2 with  $\lambda = 1.25$  and  $\mu = 1$ . We see that the estimates are highly concordant with Pearson correlation of 99.7%.

### 3.4 Small reproduction number and high sample fraction

The CoM based on a deterministic demographic process may be most biased when the population size is small and subject to large stochastic fluctuations. We generated 300 trees from the BD process with  $\lambda = 1.1, \mu = 1$  and homochronous sampling with  $n = 100$  and a variable sample fraction around 50%. The distribution of MLE birth rates is shown in Figure 6.

We found small but significant bias in the estimated birth rates using both BDM and CoM methods. The mean bias of the BDM estimator was 0.019 (95% CI: ( 0.0148, 0.0244)), and the mean bias of the CoM was 0.021 (95% CI of bias: (0.0160, 0.0265)). The BDM had smaller RMSE (0.043 vs 0.47), and the Pearson correlation of estimated birth rates was 95%. A comparison of estimated birth rates is shown in SI figure S3.

### 3.5 Decreasing sample rate and small sample fraction

When the sampling model implicit to the BDM approach is misspecified, the BDM may yield highly biased results. Figure 7 shows the MLE birth rates for both the BDM and CoM estimators when the sampling rate changes through time according to  $e^{\alpha t}$  (see section 1). 120 simulations were carried out, and the sampling rate decreased at a rate of  $\alpha = -0.44$ . This value was chosen so that the expected sample size would be 100 if taking a weighted sample of all lineages at the time of death. Note that the sampling rate is an exponential function of time, so that the sequence of sample times still appears as though it arises from an exponentially increasing population, and there would be no warning from the sequence of sample times alone that the rate is changing. The BDM estimates are biased downwards by 0.23 (95% CI: (-0.2488, -0.2207)).

In this scenario, the CoM is robust to changing sample rate, since the CoM conditions on observed sample times. The CoM estimates did not have significant bias (95% CI of bias: (-0.0443, 0.0045)).

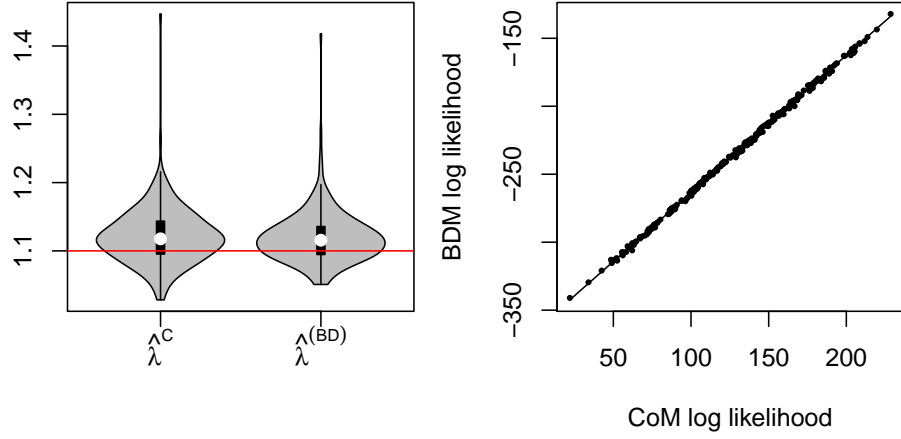


Figure 6: MLEs based on simulations with homochronous sampling,  $\lambda = 1.1, \mu = 1$  and a variable sample fraction around 50%. Left: Distribution of MLE birth rates from 300 simulations and using BDM and CoM estimators. The red line shows the true parameter value. Right: The log likelihoods of the BDM and CoM MLEs with a smoothing spline.

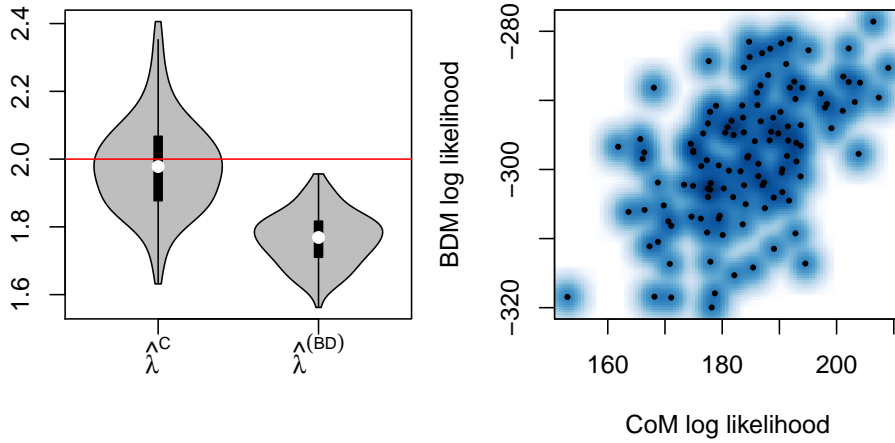


Figure 7: MLE birth rates based on simulations with time-dependent sampling. Left: Distribution of MLE birth rates from 120 simulations and using BDM and CoM estimators. Right: The log likelihoods of the BDM and CoM MLEs with a smoothing spline.

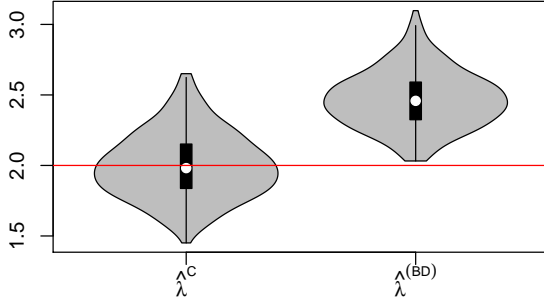


Figure 8: MLE birth rates based on simulations with a 50% sample fraction,  $n = 100$ , and with time-dependent sampling rate that increases through time. Left: Distribution of MLE birth rates from 300 simulations and using BDM and CoM estimators.

### 3.6 Increasing sample rate and large sample fraction

In these experiments, we examine bias in the coalescent due to sampling a large fraction of lineages from a small population growing stochastically. 300 genealogies with  $n = 100$  were simulated from a birth death process. Simulations were terminated when the number of deceased lineages reached 200, so that the sample fraction was 50% of deceased lineages and about 25% of all lineages. In the same experiments, we examined bias in BDMs due to a misspecified sampling process. In these experiments, the sampling rate increases from zero at time zero at a rate of  $\rho = \mu$ .

Figure 8 shows the distribution of MLE birth rates. We do not find detectable bias with the CoM estimator (95% CI: (-0.0271, 0.0260)), despite using a misspecified deterministic approximation to the demographic process, and despite that a large sample of the population was taken and that the population size was only around 400 on average at the time of the last sample.

Because the BDM relies on a misspecified sampling process, the BDM estimator gives highly biased estimates in this scenario. The average bias was 0.46 (95% CI: (0.4460, 0.4920)).

### 3.7 Asymptotic distribution of coalescent times

Some insight into why CoM and BDM give similar estimates can be gained by comparing the asymptotic distribution of coalescent times predicted by both models in the case of homochronous sampling. The distribution of coalescent times in the limit of large sample size for a deterministic coalescent model can be easily computed, and we show that this distribution is equivalent to the marginal likelihood of a node given by the birth-death model.

In [26, 27], an approximation to the lineages through time for the coalescent process was presented for a population under exponential growth:

$$\frac{d}{ds}A = -\binom{A(s)}{2} \frac{2\lambda}{Y(s)} \quad (18)$$

If sampling occurs at a single time point, such that  $A(0) = n$ , this has unique solution

$$A(s) = \frac{1}{1 - \frac{1}{n}(n-1)e^{-\frac{\lambda(e^{s(\lambda-\mu)}-1)}{Y_0(\lambda-\mu)}}}, \quad (19)$$

where  $Y_0$  is the population size at the time of sampling. We will call this a doubly deterministic coalescent model (DDCoM) because both the demographic and genealogical processes are modeled with deterministic approximations. The asymptotic distribution of coalescent times for the DDCoM is given by the derivative of  $A(s)$  (equation 18 and expanding  $Y(s)$  and normalizing:

$$P_{ddcom}(s|\lambda, \mu, \rho, n) = -\frac{d}{ds}A/(n-1) \quad (20)$$

$$= \frac{\lambda \rho e^{\frac{\lambda \rho (e^{s(\lambda-\mu)} - 1)}{n(\lambda-\mu)} + s(\lambda-\mu)}}{\left( n e^{\frac{\lambda \rho (e^{s(\lambda-\mu)} - 1)}{n(\lambda-\mu)}} - n + 1 \right)^2}. \quad (21)$$

The factor of  $n-1$  normalises the distribution since there are  $n-1$  nodes in the tree. In [28], the DDCoM was found to be an excellent approximation to the stochastic coalescent for large populations.

The BDM likelihood takes the form of a product over coalescent times and sample times, including the time of origin. Conditioning on the time of origin, and given a homochronous sample, the likelihood is given by the product of marginal probabilities for each coalescent time. From equation 6, expanding  $c_1, c_2$  and simplifying:

$$\begin{aligned} P_{bdm}(s|\lambda, \mu, \rho) &= \frac{4\lambda\rho}{q(s, c_2)} \\ &= \frac{4\lambda\rho}{2(1 - c_2^2) + e^{-c_1 s}(1 - c_2)^2 + e^{c_1 s}(1 + c_2)^2}, \end{aligned} \quad (22)$$

where  $c_1$  and  $c_2$  are the following constants:

$$c_1 = |\lambda - \mu| \quad (23)$$

$$c_2 = -(\lambda - \mu - 2\lambda\rho)/c_1 \quad (24)$$

**Theorem 3.1.** *Given a homochronous sample of a proportion  $\rho$  lineages from a population growing exponentially according to the birth-death process with birth rate  $\lambda$ , death rate  $\mu$ , and  $\lambda > \mu$ ,*

$$\lim_{n \rightarrow \infty} P_{ddcom}(s|n, \lambda, \mu, \rho) = P_{bdm}(s|\lambda, \mu, \rho)$$

for all times  $s$ .

*Proof.* By Taylor expansion of the denominator of equation 21, we have

$$\begin{aligned} &\left( n e^{\frac{\lambda \rho (e^{s(\lambda-\mu)} - 1)}{n(\lambda-\mu)}} - n + 1 \right)^2 \\ &= \left( 1 + \frac{\lambda \rho (e^{s(\lambda-\mu)} - 1)}{(\lambda - \mu)} + O(1/n) \right)^2 \end{aligned} \quad (25)$$

The limit of the numerator of equation 21 is

$$\begin{aligned} &\lim_{n \rightarrow \infty} \lambda \rho e^{\frac{\lambda \rho (e^{s(\lambda-\mu)} - 1)}{n(\lambda-\mu)} + s(\lambda-\mu)} \\ &= \lambda \rho e^{s(\lambda-\mu)} \end{aligned} \quad (26)$$

Taking the large  $n$  limit of equation 25 and computing the ratio of 25 and 26, and rearranging, we have

$$\lim_{n \rightarrow \infty} P_{ddcom}(s|\lambda, \mu, \rho) = \frac{\lambda \rho (\lambda - \mu)^2 e^{s(\lambda - \mu)}}{(\lambda - \mu - \lambda \rho + \lambda \rho e^{s(\lambda - \mu)})^2} \quad (27)$$

It may be verified that this is equivalent to  $P_{bdm}$  (equation 22).  $\square$

Note that this result applies to the DDCoM model and not the coalescent model used elsewhere in the text. In [29, 28] it was shown that the lineages through time given by DDCoMs are generally excellent approximations to lineages through time given by standard CoMs if the sample size is large.

Outside of the large- $n$  limit, we can investigate the similarity of  $P_{bdm}$  and  $P_{ddcom}$  numerically. To summarize the difference between distributions  $P_{bdm}$  and  $P_{ddcom}$ , we compute the Kullback-Leibler divergence:

$$D(P_{ddcom}, P_{bdm}|\lambda, \mu, \rho, n) = \int_{s=0}^{\infty} \log\left(\frac{P_{ddcom}(s|\lambda, \mu, \rho, n)}{P_{bdm}(s|\lambda, \mu, \rho)}\right) P_{ddcom}(s|\lambda, \mu, \rho, n) ds$$

Figure 9 shows the divergence as a function of sample sizes ranging from  $n = 2$  to  $n = 2^{14}$  and with  $\lambda = 1.1, \mu = 1$ , and  $\rho = 0.9$ . We find that divergence is very insensitive to birth rates and sample proportion, so results are only shown for one scenario. When  $n = 2$ , the divergence is quite high, but it rapidly converges to zero. We observe that to excellent approximation, the divergence scales in a very simple way as a function of sample size:  $D(P_{ddcom}, P_{bdm}|\lambda, \mu, \rho, n) \approx e^{-3/2}/n$ , and this is shown by the red line in Figure 9.

Figure 9 also shows a comparison of the DDCoM marginal density of coalescent times with the BDM marginal likelihood with several different sample sizes and a smaller sample fraction of  $\rho = 0.01$ . When  $n = 2$ , the distributions are quite different, but when  $n = 10$ , they are very similar and when  $n \geq 100$  they are almost indistinguishable.

## 4 Discussion

Two distinct areas of concern have arisen related to phylodynamic inference using coalescent models (CoMs) and birth-death-sampling models (BDMs). CoMs based on a deterministic demographic process may be subject to inductive bias if the deterministic process is a bad approximation to the true stochastic demographic process. Similarly, BDMs are subject to bias if the model of the sampling process is misspecified. We have found that the bias due to the deterministic approximation is generally very small for populations growing exponentially, even when sampling 50% of individuals from a small population. Furthermore, errors in CoMs due to a deterministic process can be resolved with additional computational effort, as it is possible to use the coalescent with a stochastic demographic process [19, 30]. Such methods may be necessary for populations with very small and noisy population dynamics. Bias is likely to be greatest if the population is small and growing slowly such that population dynamics are relatively noisy. Indeed, we found only one situation where the BDM was noticeably more precise than CoM estimators, which occurred with a small  $R_0$  of 1.1 and a large sample fraction, however, we did not find a situation where the BDM estimator was substantially less biased than the CoM estimator.

We have found that BDMs can yield highly biased estimates if the sampling process is misspecified. It may be hard to detect if the sampling process deviates from the modeled form in

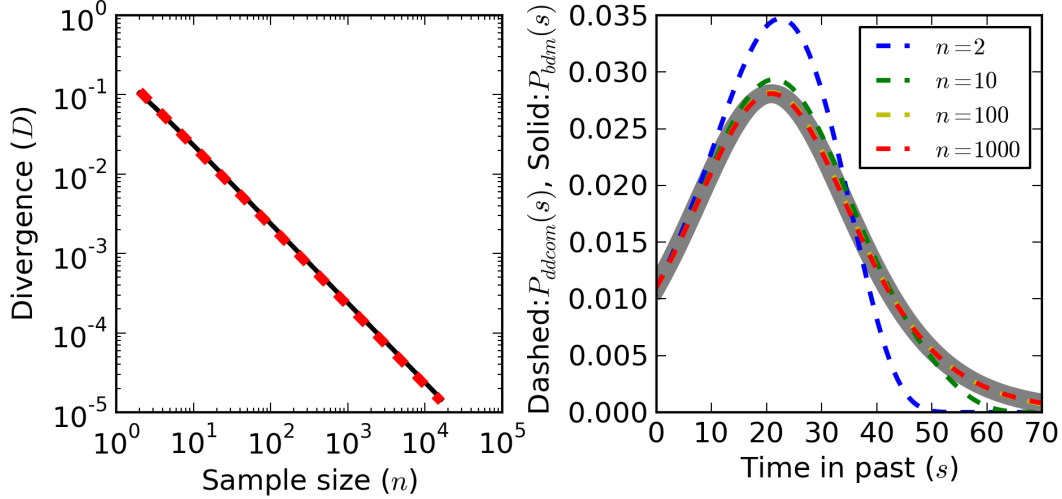


Figure 9: Left: The Kullback-Leibler divergence between the coalescent and birth-death distribution (black line) is shown versus the sample size on logarithmic axes. The red line shows a linear approximation ( $e^{-3/2}/n$ ).  $\lambda = 1.1, \mu = 1$  and  $\rho = .9$ . Right: The birth-death marginal density of a node (grey) is compared with the coalescent density based on samples of size  $n = 2, 10, 100$ , and 1000.  $\lambda = 1.1, \mu = 1$  and  $\rho = 0.01$ .

many real world situations, and most real datasets are likely to violate the BDM sampling process assumptions to some degree. An example heterogeneous sampling through time is shown in Figure 10 for a dataset which has previously been analyzed with BDMs in [31]. Figure 10 shows the sampling proportion through time of HIV sequence samples in the UK Resistance Database [32]. Typical for HIV sequence databases, the sample proportion is essentially zero throughout the 80s, and there is a rapid increase in sampling effort throughout the late 90s and early 00s, followed by a plateau after 2010 due to reporting delays. In [31] a BD SIR model was fitted to HIV sequence data from the UK under the assumption of a constant sampling rate, but the timespan of the estimated phylogenies ranged from 1978-2003 over which the true sampling rate varied greatly. Future work should explore how violation of sampling assumptions may bias estimates of  $R_0$  when fitting BDM SIR models.

The sequence of sample times may be informative about the population size through time if the sampling process can be correctly specified. We have shown how birth rates may be estimated from the sequence of sample times if sampling occurs according to the BDM assumptions, and this is possible even if the sample rate is not known. BDMs implicitly use the sequence of sample times to estimate birth and/or death rates, and this is the case even if the sampling rate is not given, but estimated. Comparisons of CoMs and BDMs should account for the effects of sampling, and a fair comparison can be obtained in the case of homochronous or serial-homochronous sampling with unknown sample rate, so that the sample times contain no information about population size and birth rates.

Previous simulation-based studies on fitting susceptible-infected-recovered (SIR) epidemiological models to sequence data [30] have purported to show increased statistical efficiency of BDMs relative to CoMs, but these studies did not control for the informativeness of sample times, and the supposed advantage of BDM in these simulations is likely due to the sampling model and not the genealogical model. For example, the simulation studies in [30] did not consider a homochronous



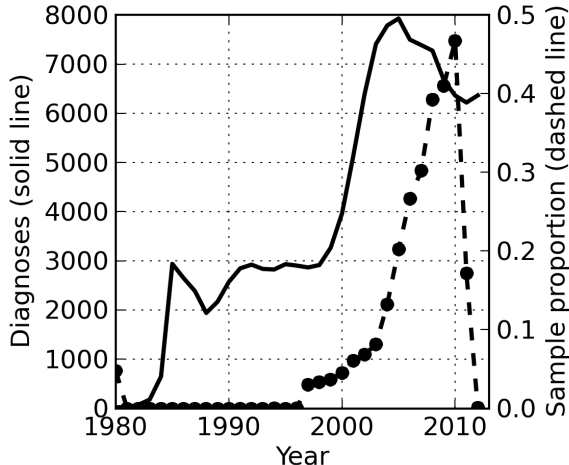


Figure 10: The HIV sequence sample rate through time using data from the HIV resistance database [32] and the number living diagnosed with HIV through time.

sample, a misspecified sampling process, or the possibility of extending the coalescent estimators to use sample time information. The study in [30] used a Bayesian method, in contrast to our ML methods, so some differences may also be due to the choice of priors. It was claimed in [30] that the difference in performance of BDMs and CoMs was due to the latter’s use of a misspecified deterministic demographic process, but in the context of exponential growth, we found very little bias due to the deterministic approximations of the coalescent, but large biases due to the effects of sampling.

Future research on BDMs may reveal ways to accomodate more realistic sampling processes. For example, in [22], a piecewise constant sampling process was presented, however this also required the introduction of many more parameters to describe the sampling process. If the sampling process is known, a useful alternative to BDMs is model the sampling process in tandem with the coalescent. As we have shown, the coalescent likelihood of a genealogy is approximately independent of the likelihood of the sample times, and for complex sampling processes it is much easier to model the genealogical and sampling process separately and combine likelihoods then to derive a joint likelihood. In most cases, the sampling process is unknown, but we have shown that CoMs are robust to a diverse range of sampling scenarios.

**Acknowledgements.** The authors thank Alison Brown (PHE UK) for providing HIV diagnosis statistics from the UK. Xavier Didelot and Caroline Colijn (Imperial College London) provided many helpful suggestions. SDWF is supported in part by an MRC Methodology Research Programme grant (MR/J013862/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## References

- [1] Volz EM, Koelle K, Bedford T. Viral phylodynamics. *PLoS computational biology*. 2013;9(3):e1002947.
- [2] Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole genome sequence data. *Molecular biology and evolution*. 2014;p. msu121.

- [3] Ypma R, Bataille A, Stegeman A, Koch G, Wallinga J, Van Ballegooijen W. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B: Biological Sciences*. 2011;p. rspb20110913.
- [4] Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, et al. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Science translational medicine*. 2012;4(148):148ra116–148ra116.
- [5] Kingman JFC. The coalescent. *Stochastic processes and their applications*. 1982;13(3):235–248.
- [6] Wakeley J. *Coalescent theory: an introduction*. vol. 1. Roberts & Company Publishers Greenwood Village, Colorado; 2009.
- [7] Rannala B, Yang Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*. 1996;43(3):304–311.
- [8] Kendall DG. On the generalized” birth-and-death” process. *The annals of mathematical statistics*. 1948;p. 1–15.
- [9] Stadler T. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*. 2010;267(3):396–404.
- [10] Rosenberg NA, Nordborg M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*. 2002;3(5):380–390.
- [11] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*. 2010;59(3):307–321.
- [12] Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*. 2012;29(8):1969–1973.
- [13] Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*. 2000 Jul;155(3):1429–1437.
- [14] Strimmer K, Pybus OG. Exploring the demographic history of DNA sequences using the generalized e plot. *Molecular Biology and Evolution*. 2001;18(12):2298–2305.
- [15] Opgen-Rhein R, Fahrmeir L, Strimmer K. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC evolutionary biology*. 2005;5(1):6.
- [16] Donnelly P, Tavaré S. Coalescents and genealogical structure under neutrality. *Annual review of genetics*. 1995;29(1):401–421.
- [17] Volz EM. Complex population dynamics and the coalescent under neutrality. *Genetics*. 2012;190(1):187–201.
- [18] Rasmussen DA, Ratmann O, Koelle K. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS computational biology*. 2011;7(8):e1002136.

- [19] Rasmussen DA, Volz EM, Koelle K. Phylodynamic Inference for Structured Epidemiological Models. *PLoS computational biology*. 2014;10(4):e1003570.
- [20] Thompson EA. Human evolutionary trees. CUP Archive; 1975.
- [21] Gernhard T. The conditioned reconstructed process. *Journal of theoretical biology*. 2008;253(4):769–778.
- [22] Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*. 2013;110(1):228–233.
- [23] Stadler T, Bonhoeffer S. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2013;368(1614):20120198.
- [24] Vaughan TG, Drummond AJ. A Stochastic Simulator of Birth–Death Master Equations with Application to Phylodynamics. *Molecular biology and evolution*. 2013;30(6):1480–1493.
- [25] Leventhal GE, Günthard HF, Bonhoeffer S, Stadler T. Using an Epidemiological Model for Phylogenetic Inference Reveals Density Dependence in HIV Transmission. *Molecular biology and evolution*. 2014;31(1):6–17.
- [26] Frost SD, Volz EM. Viral phylodynamics and the search for an effective number of infections. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2010;365(1548):1879–1890.
- [27] Maruvka YE, Shnerb NM, Bar-Yam Y, Wakeley J. Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate. *Molecular biology and evolution*. 2011;28(5):1617–1631.
- [28] Jewett EM, Rosenberg NA. Theory and applications of a deterministic approximation to the coalescent model. *Theoretical population biology*. 2014;93:14–29.
- [29] Chen H, Chen K. Asymptotic Distributions of Coalescence Times and Ancestral Lineage Numbers for Populations with Temporally Varying Size. *Genetics*. 2013;194(3):721–736.
- [30] Poppinga A, Vaughan T, Stadler T, Drummond A. Bayesian Coalescent Epidemic Inference: Comparison of Stochastic and Deterministic SIR Population Dynamics. *arXiv preprint arXiv:14071792*. 2014;.
- [31] Kühnert D, Stadler T, Vaughan TG, Drummond AJ. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *Journal of The Royal Society Interface*. 2014;11(94):20131106.
- [32] Brown AJL, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *Journal of Infectious Diseases*. 2011;p. jir550.

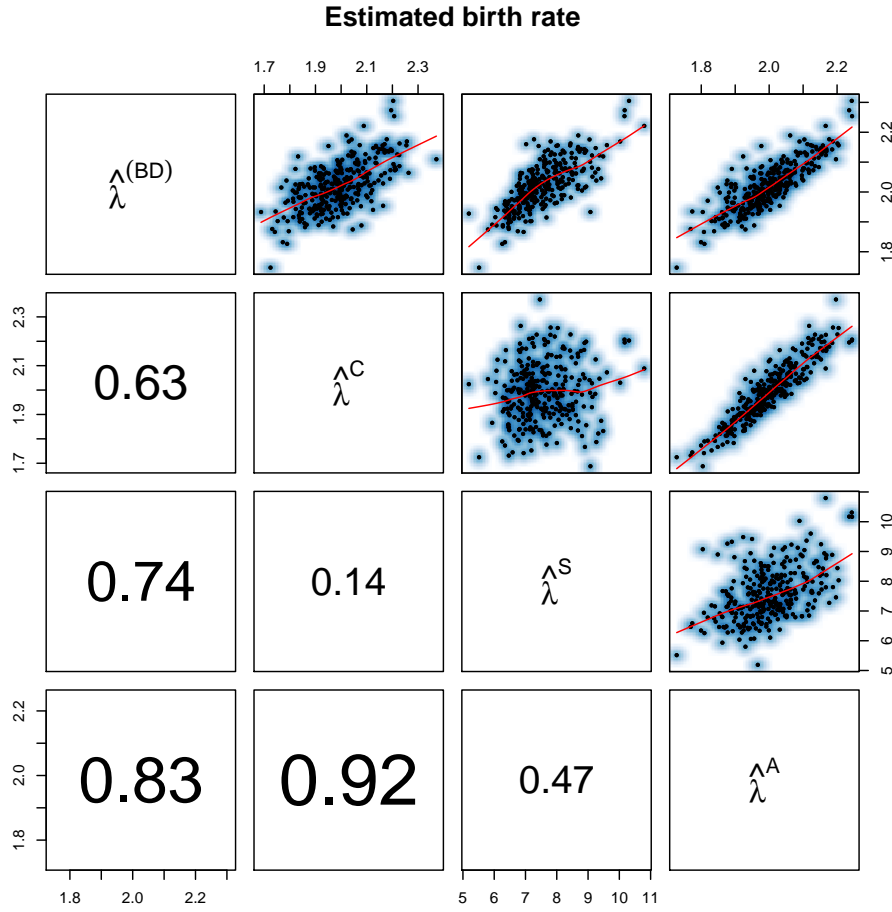


Figure S1: The MLE birth rate using four estimation methods and based on 300 simulated genealogies. The Pearson correlation coefficient between the estimates is shown in the lower panels. The upper panels show a scatter plot with smoothing splines.

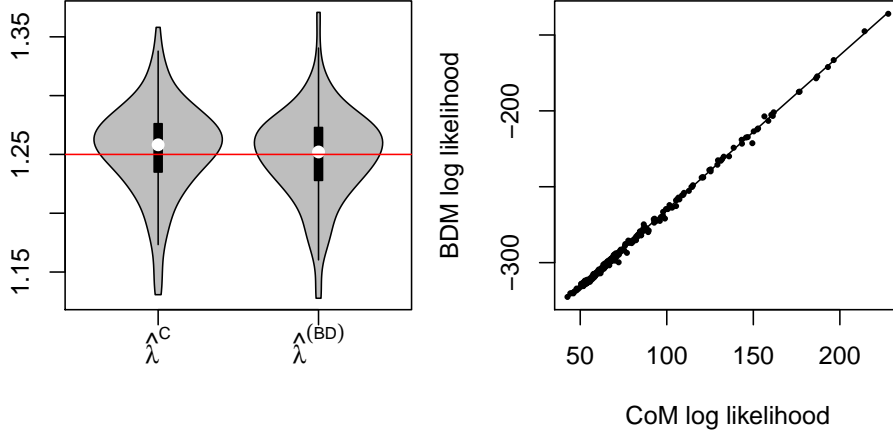


Figure S2: MLEs based on simulations with homochronous sampling and  $\lambda = 1.25$ . Left: Distribution of MLE birth rates from 300 simulations and using BDM and CoM estimators. The red line shows the true parameter value. Right: The log likelihoods of the BDM and CoM MLEs with a smoothing spline.

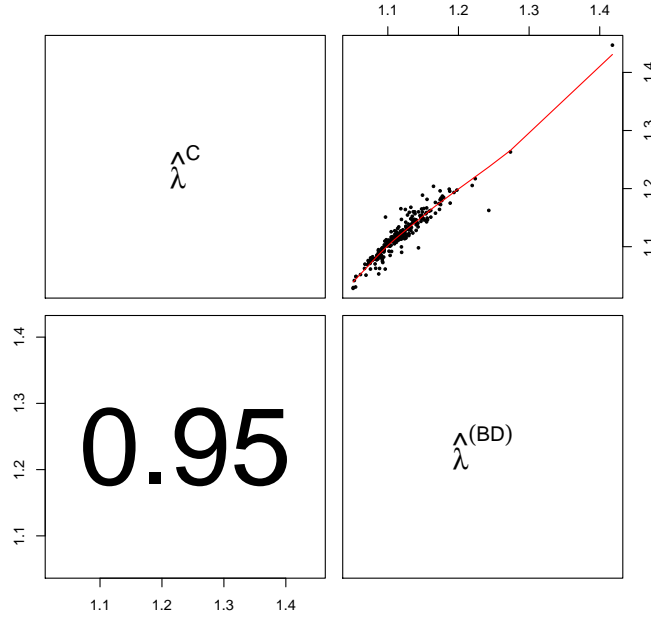


Figure S3: The BDM and CoM MLE birth rate based on 300 simulated genealogies with  $\lambda = 1.1$  and large sample fraction. The Pearson correlation coefficient between the estimates is shown in the lower panels. The upper panels show a scatter plot with smoothing splines.